Analyzing NOAA Storm Damage Data

MATTHEW TUTTLE & LEX BUKOWSKI

Project Description

- NOAA Provides data on storms that have caused injuries or significant property damage dating back to 1950. The data set describes the location of event, type of weather, and magnitude of damage. There are approximately 1.8 MM weather events recorded.
- Prior Work:
  - “Student Research Abstract: Unsupervised Key Term Extraction of Tornado Narratives from NOAA Storm Events Database” –Emma Louise McDaniel
    - Narratives were text mined in order to retrieve the impacts of the disasters in order to structure the data for further use
  - “Some Comments on the Reliability of NOAA's Storm Events Database” –Renato P Dos Santos
    - Supported by limited statistical analysis, came to the conclusion that the database suffers from incompleteness and inconsistencies
- Datasets:
  - https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/

Questions

- Have weather events become more dangerous to people?
- Is the magnitude of property damage indicative of higher risk of injury to people?
- Has the damage caused by severe weather events increased over time?
- Have severe weather events increased over time?
- Are some areas more prone to severe weather effects?

Data Preparation

- Data Cleaning & Preprocessing
  - Data broken up into yearly reports which were combined in Pandas dataframe
  - Simplified, combined and made data types consistent between files

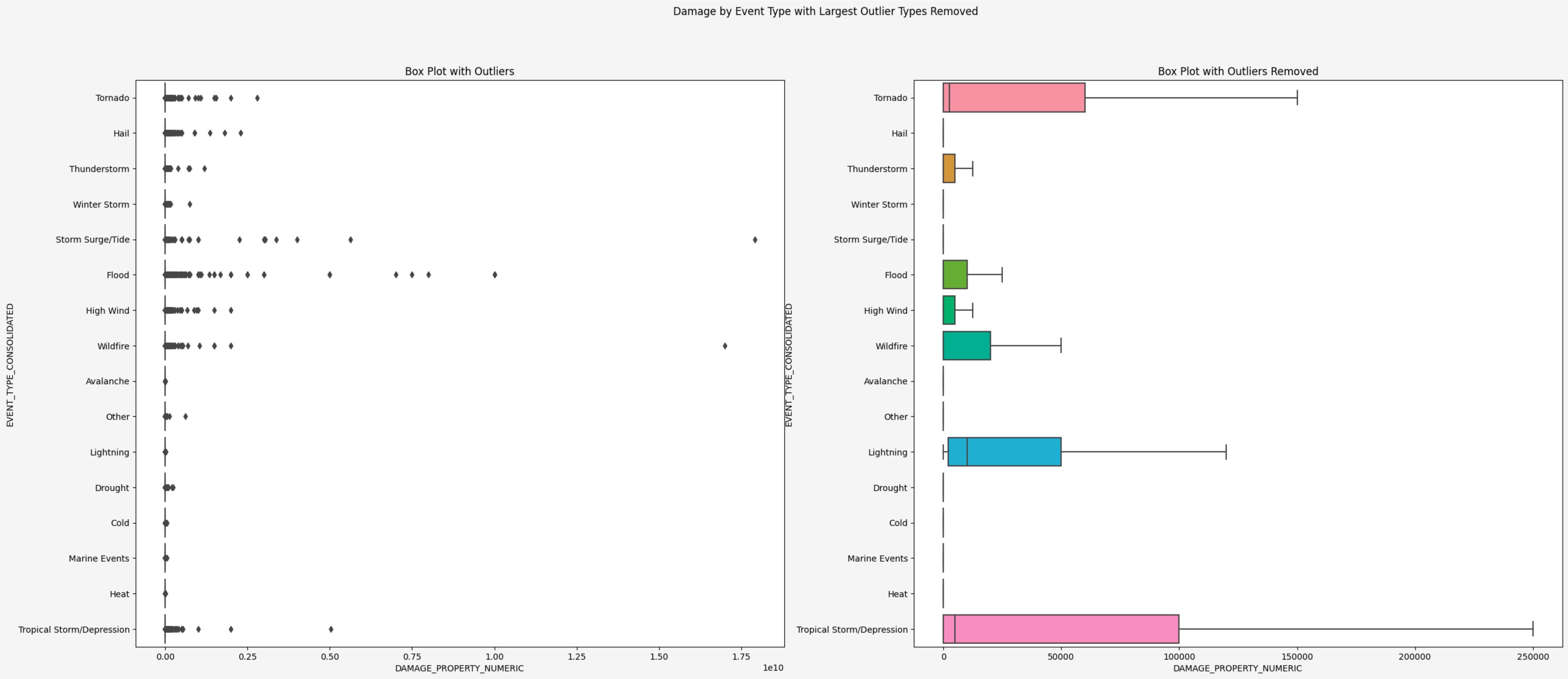
- Data Integration
  - Codes for county and regional data were re-labeled for readability and consistency
  - Census data per county

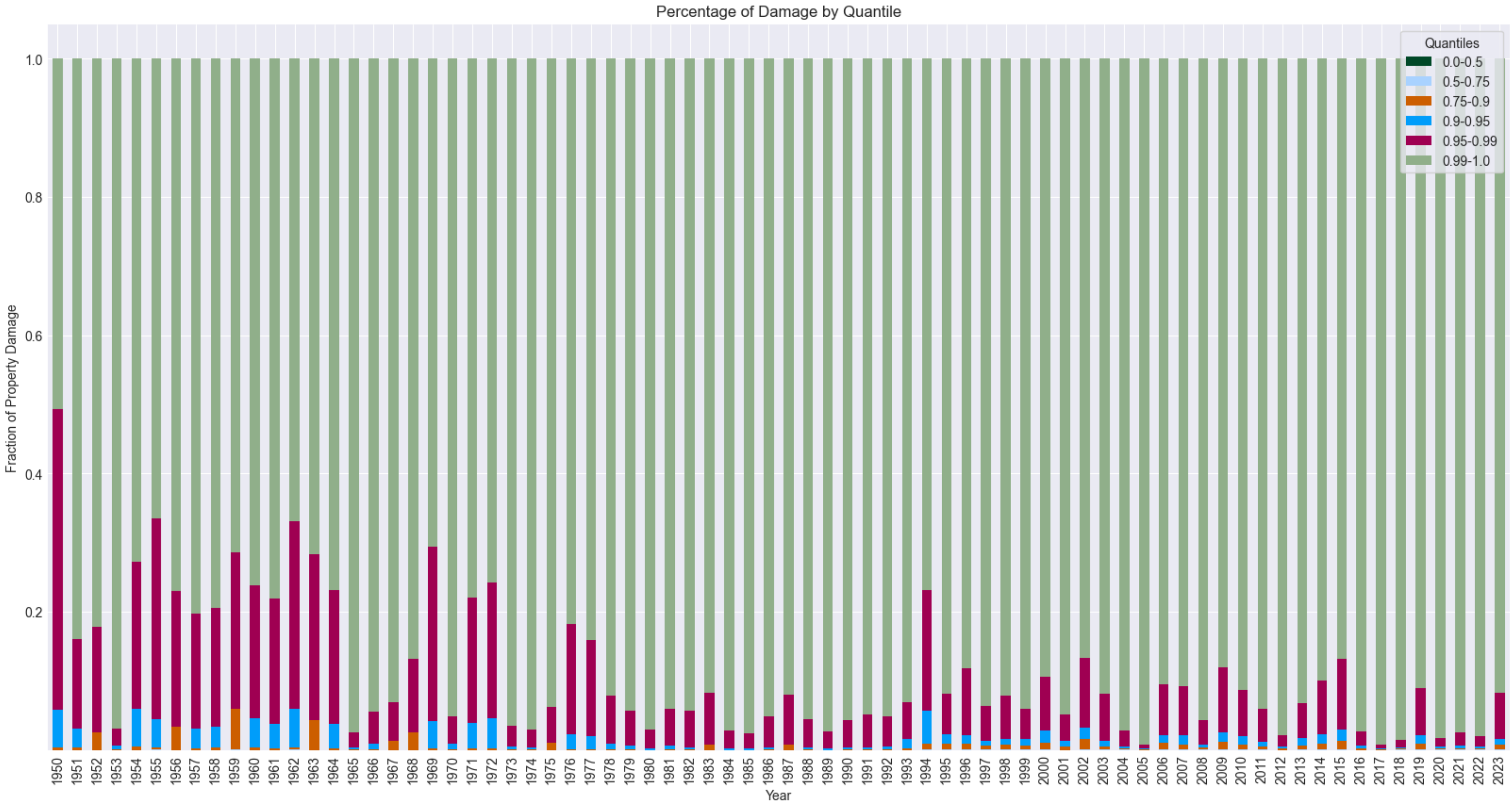
Tools Used

- Pandas
  - Created a Pandas Dataframe in Python from the CSV Storm Event Data Files
  - Allowed for efficient data cleaning and preliminary analysis
- NumPy
  - Python tool to help clean, pre-process, and simplify the data set
- StatsModels
  - Statistical modeling and regression calculations for correlation analysis to answer the proposed questions
- Seaborn and Plotly
  - Tool to create graphs and data visualizations

Data Mining Analysis Applied

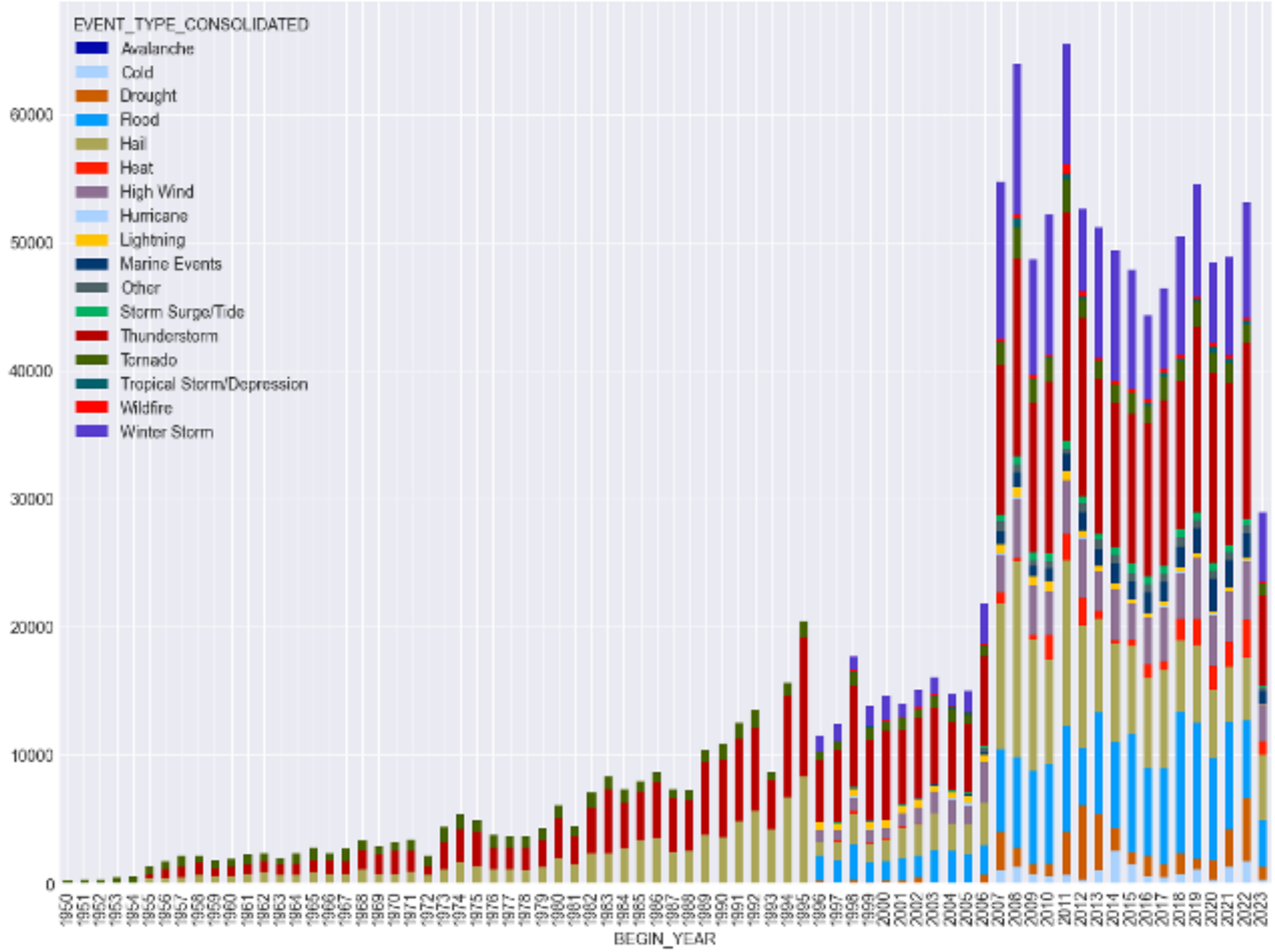
- Regression Models
  - Determine relevant correlations between data attributes to predict future storm event damage cost and injury counts
- Outlier Analysis
  - Normal EDA tools such as histogram and box plots were not effective due to dominance of outliers in dataset


Damage by Event Type with Largest Outlier Types Removed
Box Plot with Outliers
EVENT_TYPE_CONSOLIDATED
Tornado
Hail
Thunderstorm
Winter Storm
Storm Surge/Tide
Flood
High Wind
Wildfire
Avalanche
Other
Lightning
Drought
Cold
Marine Events
Heat
Tropical Storm/Depression
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e10
DAMAGE_PROPERTY_NUMERIC
Box Plot with Outliers Removed
EVENT_TYPE_CONSOLIDATED
Tornado
Hail
Thunderstorm
Winter Storm
Storm Surge/Tide
Flood
High Wind
Wildfire
Avalanche
Other
Lightning
Drought
Cold
Marine Events
Heat
Tropical Storm/Depression
0
50000
100000
150000
200000
250000
DAMAGE_PROPERTY_NUMERIC


Percentage of Damage by Quantile
Fraction of Property Damage
Year
0.2
0.4
0.6
0.8
1.0
Quantiles
0.0-0.5
0.5-0.75
0.75-0.9
0.9-0.95
0.95-0.99
0.99-1.0
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

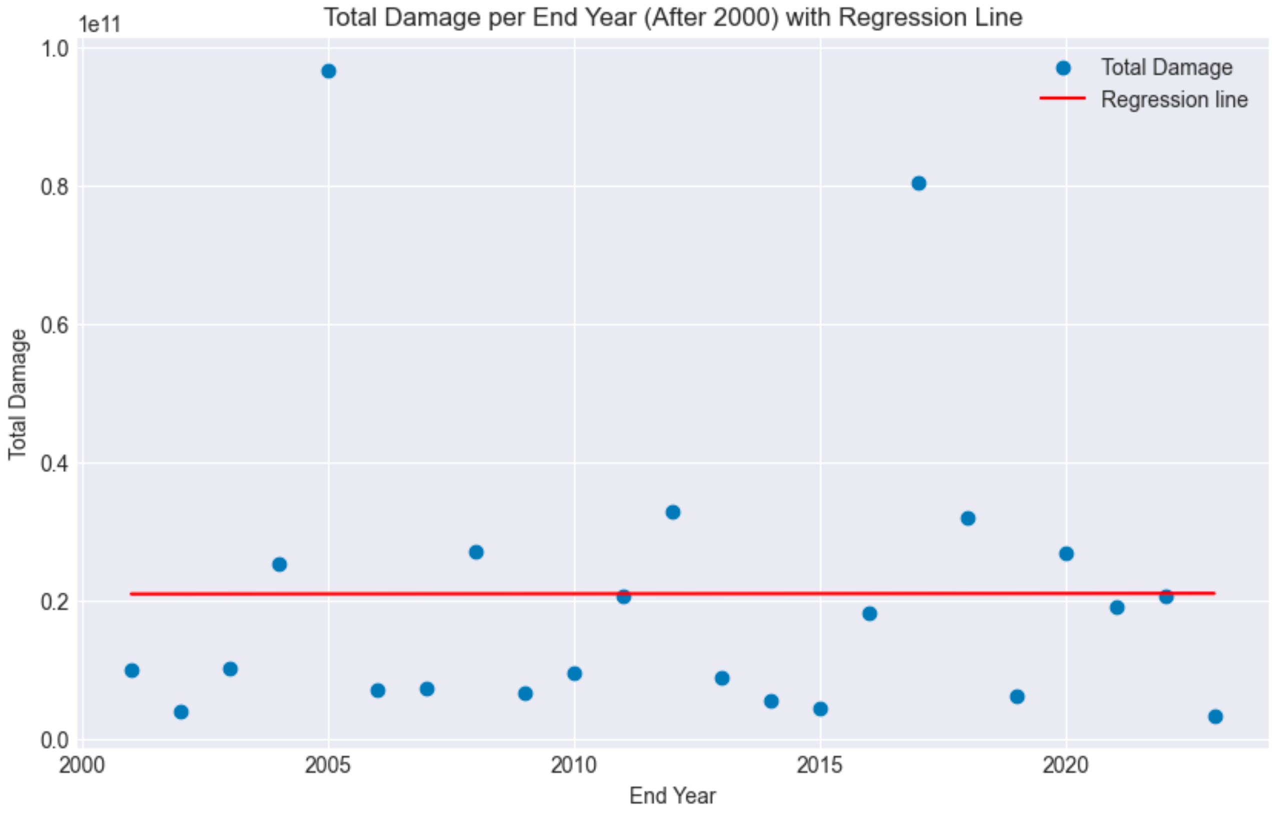
Knowledge Gained – Events per Year

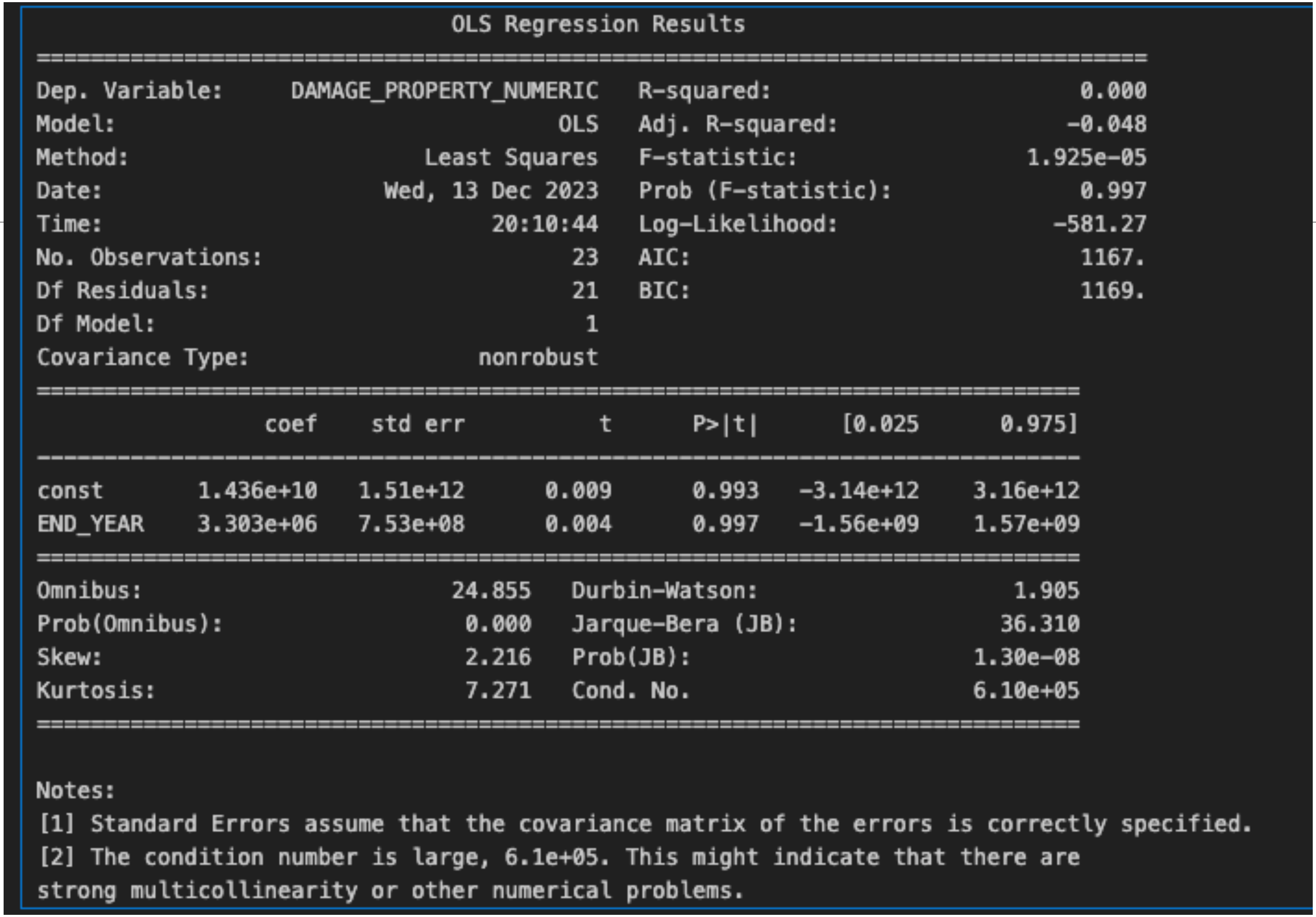
- Has the frequency of storm events increased over time?
- No trend to show an increase in total weather events each year
- Recording of event types has changed significantly in the lifetime of dataset



Knowledge Gained – Damage per Year

- Has the damage caused by storm events increased over time?
- R-squared value = 0.000
- No correlation



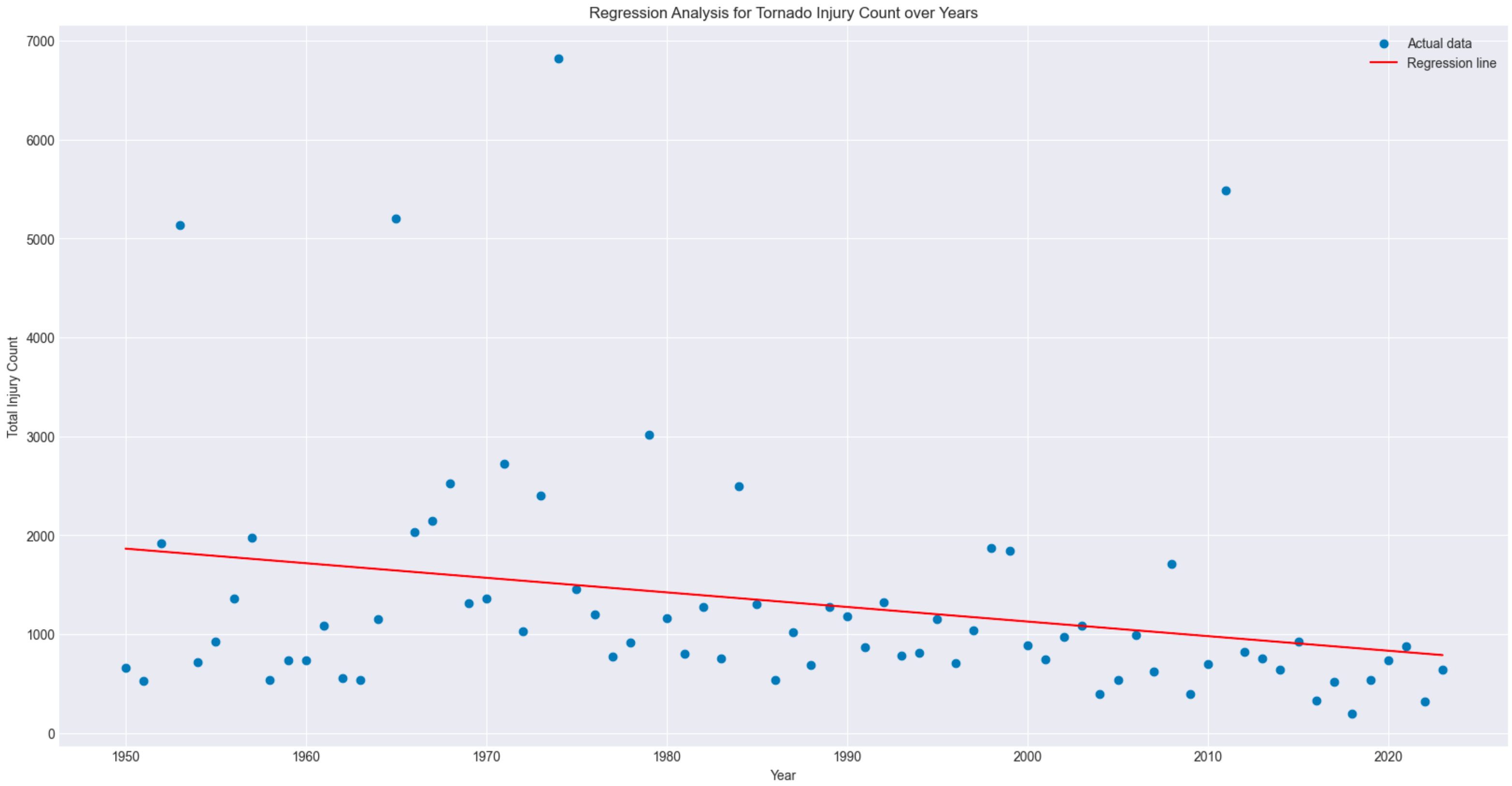


```
                               OLS Regression Results
===================================================================================
Dep. Variable:     DAMAGE_PROPERTY_NUMERIC   R-squared:                       0.000
Model:                                 OLS   Adj. R-squared:                 -0.048
Method:                      Least Squares   F-statistic:                 1.925e-05
Date:                     Wed, 13 Dec 2023   Prob (F-statistic):              0.997
Time:                             20:10:44   Log-Likelihood:                -581.27
No. Observations:                       23   AIC:                             1167.
Df Residuals:                           21   BIC:                             1169.
Df Model:                                1
Covariance Type:                 nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.436e+10   1.51e+12      0.009      0.993   -3.14e+12    3.16e+12
END_YEAR    3.303e+06   7.53e+08      0.004      0.997   -1.56e+09    1.57e+09
==============================================================================
Omnibus:                       24.855   Durbin-Watson:                   1.905
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.310
Skew:                           2.216   Prob(JB):                     1.30e-08
Kurtosis:                       7.271   Cond. No.                     6.10e+05
==============================================================================


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.1e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

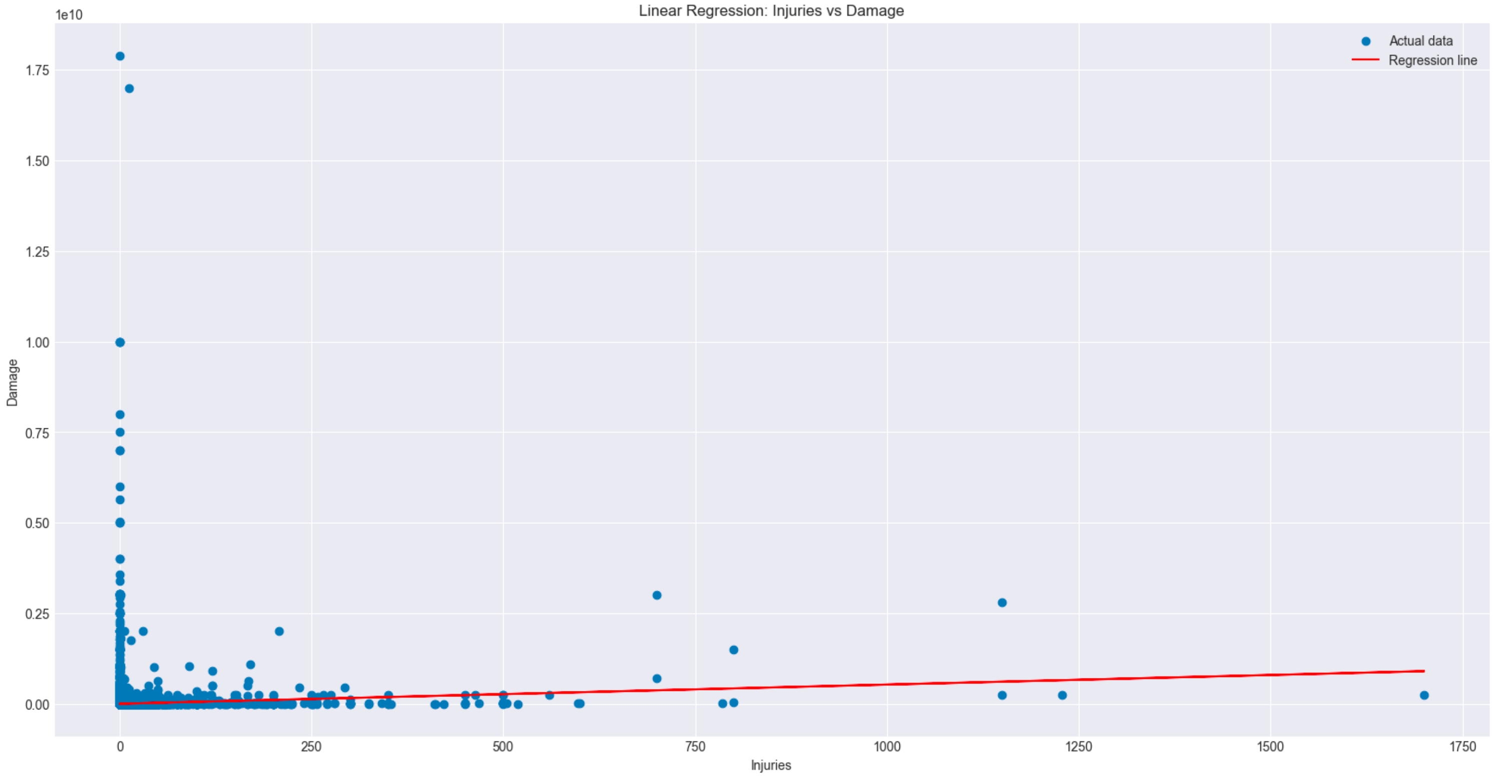
Knowledge Gained – Injury Count

- Have weather events become more dangerous to people?
- R-squared = 0.068
- Outliers skew regression model

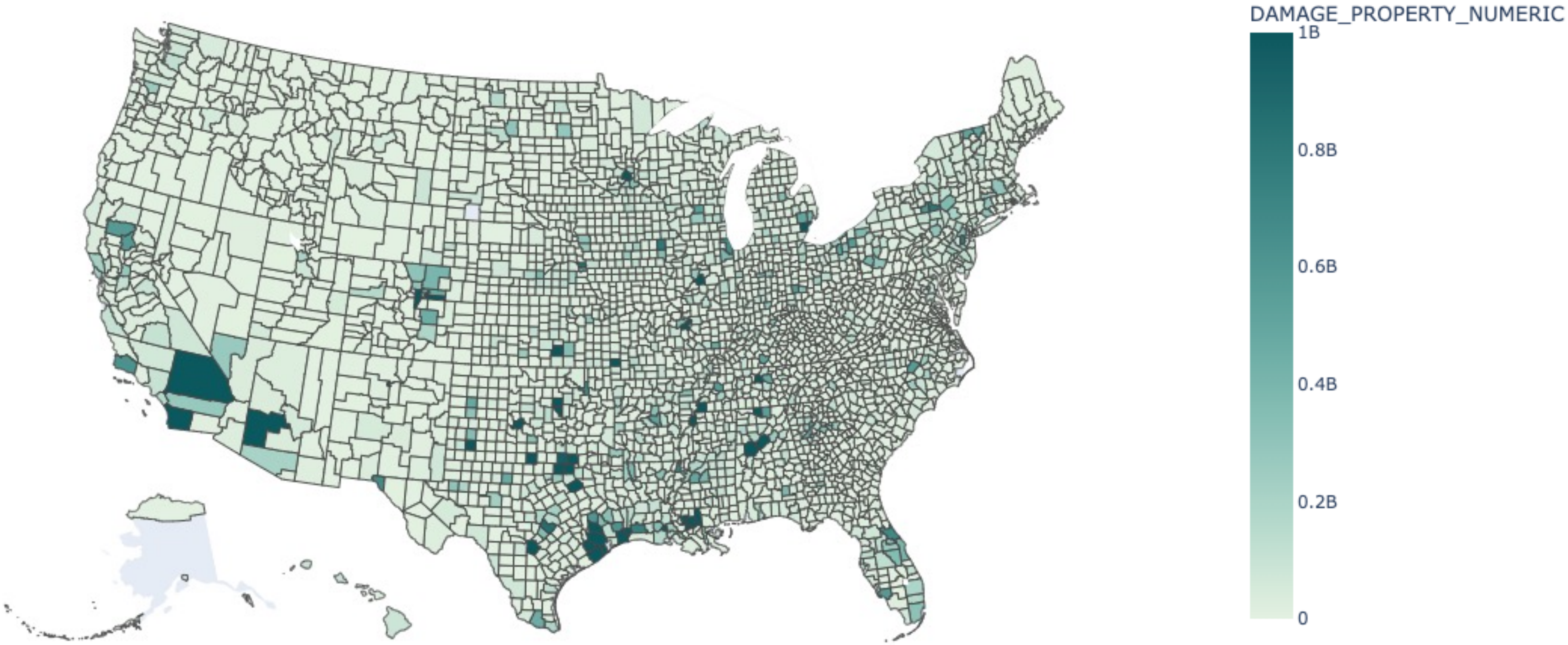


Knowledge Gained – Injuries vs Damage

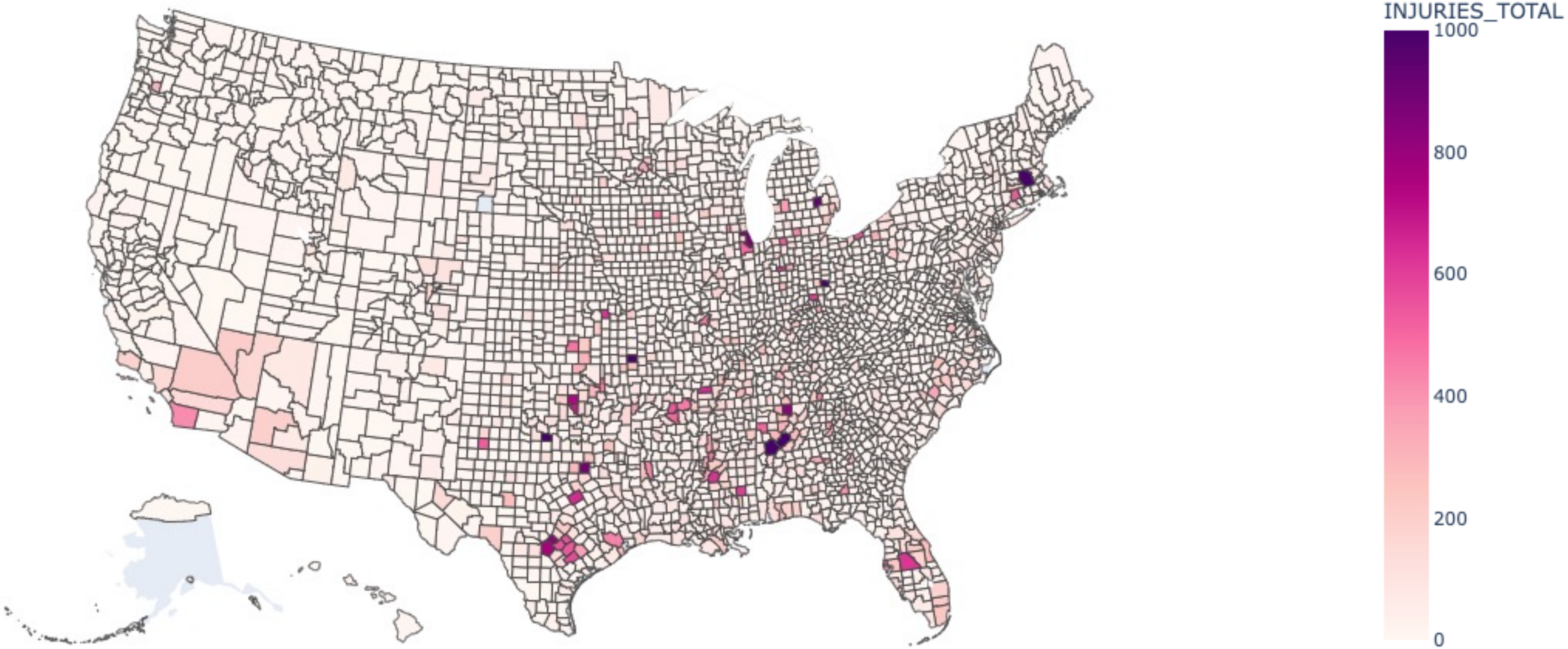
- Is the magnitude property damage indicative of higher risk of injury to people?
- R-squared value = 0.004, no significant correlation



Total Property Damage by County



Total Injuries by County



Knowledge Application

- Better understanding of the risk of severe weather events and allow for better prediction of the outcomes of severe weather events.
    - Future weather events, even if more severe, can cause less damage to people and property, potentially saving lives and preventing huge expenditures.
- Occurrence and location of weather events can give insight into possible other causation factors

Acknowledgements

Louise McDaniel and Renato P Dos Santos for their previous work and insights into this data set

SAC '23: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. March 2023 Pages 653 – 656 https://doi.org/10.1145/3555776.3577211

P dos Santos, Renato, Some Comments on the Reliability of NOAA's Storm Events Database (June 22, 2016). Available at SSRN: https://ssrn.com/abstract=2799273 or http://dx.doi.org/10.2139/ssrn.2799273